

# Semantic Tagging with Deep Residual Networks

**Johannes Bjerva**

University of Groningen  
The Netherlands  
j.bjerva@rug.nl

**Barbara Plank**

University of Groningen  
The Netherlands  
b.plank@rug.nl

**Johan Bos**

University of Groningen  
The Netherlands  
johan.bos@rug.nl

## Abstract

We propose a novel semantic tagging task, *semtagging*, tailored for the purpose of multilingual semantic parsing, and present the first tagger using deep residual networks (ResNets). Our tagger uses both word and character representations and includes a novel residual bypass architecture. We evaluate the tagset both intrinsically on the new task of semantic tagging, as well as on Part-of-Speech (POS) tagging. Our system, consisting of a ResNet and an auxiliary loss function predicting our semantic tags, significantly outperforms prior results on English Universal Dependencies POS tagging (95.71% accuracy on UD v1.2 and 95.67% accuracy on UD v1.3).

## 1 Introduction

A key issue in computational semantics is the transferability of semantic information across languages. Many semantic parsing systems depend on sources of information such as POS tags (Pradhan et al., 2004; Copestake et al., 2005; Bos, 2008; Butler, 2010; Berant and Liang, 2014). However, these tags are often customised for the language at hand (Marcus et al., 1993) or massively abstracted, such as the Universal Dependencies tagset (Nivre et al., 2016). Furthermore, POS tags are syntactically oriented, and therefore often contain both irrelevant and insufficient information for semantic analysis and deeper semantic processing. This means that, although POS tags are highly useful for many downstream tasks, they are unsuitable both for semantic parsing in general, and for tasks such as recognising textual entailment.

We present a novel set of semantic labels tailored for the purpose of multilingual semantic parsing. This tagset (i) abstracts over POS and named entity types; (ii) fills gaps in semantic modelling by adding new categories (for instance for phenomena like negation, modality, and quantification); and (iii) generalises over specific languages (see Section 2). We introduce and motivate this new task in this paper, and refer to it as *semantic tagging*. Our experiments aim to answer the following two research questions:

1. Given an annotated corpus of semantic tags, it is straightforward to apply off-the-shelf sequence taggers. Can we significantly outperform these with recent neural network architectures?
2. Semantic tagging is essential for deep semantic parsing. Can we find evidence that semtags are effective also for other NLP tasks?

To address the first question, we will look at convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are both highly prominent approaches in recent natural language processing (NLP) literature. A recent development is the emergence of deep residual networks (ResNets), a building block for CNNs. ResNets consist of several stacked residual units, which can be thought of as a collection of convolutional layers coupled with a ‘shortcut’ which aids the propagation of the signal in a neural network. This allows for the construction of much deeper networks, since keeping a ‘clean’ information path in the network facilitates optimisation (He et al., 2016). ResNets have recently shown state-of-the-art performance for image classification tasks (He et al., 2015; He et al., 2016), and have

also seen some recent use in NLP (Östling, 2016; Conneau et al., 2016). However, no previous work has attempted to apply ResNets to NLP tagging tasks.

To answer our second question, we carry out an extrinsic evaluation exercise. We investigate the effect of using semantic tags as an auxiliary loss for POS tagging. Since POS tags are useful for many NLP tasks, it follows that semantic tags must be useful if they can improve POS tagging.

## 2 Semantic Tagging

### 2.1 Background

We refer to *semantic tagging*, or *semtagging*, as the task of assigning semantic class categories to the smallest meaningful units in a sentence. In the context of this paper these units can be morphemes, words, or multi-word expressions. The linguistic information traditionally obtained for deep processing is insufficient for fine-grained lexical semantic analysis. The widely used Penn Treebank (PTB) Part-of-Speech tagset (Marcus et al., 1993) does not make the necessary semantic distinctions, in addition to containing redundant information for semantic processing. Let us consider a couple of examples.

There are significant differences in meaning between the determiners *every* (universal quantification), *no* (negation), and *some* (existential quantification), but they all receive the DT (determiner) POS label in PTB. Since determiners form a closed class, one could enumerate all word forms for each class. Indeed some recent implementations of semantic parsing follow this strategy (Bos, 2008; Butler, 2010). This might work for a single language, but it falls short when considering a multilingual setting. Furthermore, determiners like *any* can have several interpretations and need to be disambiguated in context.

Semantic tagging does not only apply to determiners, but reaches all parts of speech. Other examples where semantic classes disambiguate are reflexive versus emphasising pronouns (both POS tagged as PRP, personal pronoun); the comma, that could be a conjunction, disjunction, or apposition; intersective vs. subsecutive and privative adjectives (all POS tagged as JJ, adjective); proximal vs. medial and distal demonstratives (see Example 1); subordinate vs. coordinate discourse relations; agent nouns vs. entity nouns. The set of semantic tags that we use in this paper is established in a data-driven manner, considering four languages in a parallel corpus (English, German, Dutch and Italian). This first inventory of classes comprises 13 coarse-grained tags and 66 fine-grained tags (see Table 1). As can be seen from this table and the examples given below, the tagset also includes named entity classes (see also Example 2).

(1) *These cats live in that house .*

PRX CON ENS REL DST CON NIL

(2) *Ukraine 's glory has not yet perished , neither her freedom .*

GPE HAS CON ENT NOT IST EXT NIL NOT HAS CON NIL

In Example 1, both *these* and *that* would be tagged as DT. However, with our semantic tagset, they are disambiguated as PRX (proximal) and DST (distal). In Example 2, *Ukraine* is tagged as GPE rather than NNP.

### 2.2 Annotated Data

We use two semtag datasets. The Groningen Meaning Bank (GMB) corpus of English texts (1.4 million words) containing silver standard semantic tags obtained by running a simple rule-based semantic tagger (Bos et al., Forthcoming). This tagger uses POS and named entity tags available in the GMB (automatically obtained with the C&C tools (Curran et al., 2007) and then manually corrected), as well as a set of manually crafted rules to output semantic tags. Some tags related to specific phenomena were hand-corrected in a second stage.

Our second dataset is smaller but equipped with gold standard semantic tags and used for testing (PMB, the Parallel Meaning Bank). It comprises a selection of 400 sentences of the English part of a parallel corpus. It has no overlap with the GMB corpus. For this dataset, we used the Elephant tokeniser, which performs word, multi-word and sentence segmentation (Evang et al., 2013). We then used the simple rule-based semantic tagger described above to get an initial set of tags. These tags were then corrected by a human annotator (one of the authors of this paper).

ANA PRO pronoun	COM EQA equative	EVE EXS untensed simple
DEF definite	MOR comparative pos.	ENS present simple
HAS possessive	LES comparative neg.	EPS past simple
REF reflexive	TOP pos. superlative	EFS future simple
EMP emphasizing	BOT neg. superlative	EXG untensed prog.
ACT GRE greeting	ORD ordinal	ENG present prog.
ITJ interjection	DEM PRX proximal	EPG past prog.
HES hesitation	MED medial	EFG future prog.
QUE interrogative	DST distal	EXT untensed perfect
ATT QUA quantity	DIS SUB subordinate	ENT present perfect
UOM measurement	COO coordinate	EPT past perfect
IST intersective	APP appositional	EFT future perfect
REL relation	MOD NOT negation	ETG perfect prog.
RLI rel. inv. scope	NEC necessity	ETV perfect passive
SST subsective	POS possibility	EXV passive
PRI privative	ENT CON concept	TNS NOW present tense
INT intensifier	ROL role	PST past tense
SCO score	NAM GPE geo-political ent.	FUT future tense
LOG ALT alternative	PER person	TIM DOM day of month
EXC exclusive	LOC location	YOC year of century
NIL empty	ORG organisation	DOW day of week
DIS disjunct./exist.	ART artifact	MOY month of year
IMP implication	NAT natural obj./phen.	DEC decade
AND conjunct./univ.	HAP happening	CLO clocktime
BUT contrast	URL url	

Table 1: Semantic Tagset

For the extrinsic evaluation, we use the English portion of the Universal Dependencies dataset, version 1.2 and 1.3 (Nivre et al., 2016). An overview of the data used is shown in Table 2.

CORPUS	TRAIN (SENTS/TOKS)	DEV (SENTS/TOKS)	TEST (SENTS/TOKS)	N TAGS
ST Silver (GMB)	42,599 / 930,201	6,084 / 131,337	12,168 / 263,516	66
ST Gold (PMB)	n/a	n/a	356 / 1,718	66
UD v1.2 / v1.3	12,543 / 204,586	2,002 / 25,148	2,077 / 25,096	17

Table 2: Overview of the semantic tagging data (ST) and universal dependencies (UD) data.

### 3 Method

Our tagger is a hierarchical deep neural network consisting of a bidirectional Gated Recurrent Unit (GRU) network at the upper level, and a Convolutional Neural Network (CNN) and/or Deep Residual Network (ResNet) at the lower level, including an optional novel residual bypass function (cf. Figure 1).

#### 3.1 Gated Recurrent Unit networks

GRUs (Cho et al., 2014) are a recently introduced variant of RNNs, and are designed to prevent vanishing gradients, thus being able to cope with longer input sequences than vanilla RNNs. GRUs are similar to the more commonly-used Long Short-Term Memory networks (LSTMs), both in purpose and implementation (Chung et al., 2014). A bi-directional GRU is a GRU which makes both forward and backward passes over sequences, and can therefore use both preceding and succeeding contexts to predict a tag (Graves and Schmidhuber, 2005; Goldberg, 2015). Bi-directional GRUs and LSTMs have been shown to yield high performance on several NLP tasks, such as POS tagging, named entity tagging, and chunking (Wang et al., 2015; Yang et al., 2016; Plank et al., 2016). We build on previous approaches by combining bi-GRUs with character representations from a basic CNN and ResNets.

#### 3.2 Deep Residual Networks

Deep Residual Networks (ResNets) are built up by stacking residual units. A residual unit can be expressed as:

$$\begin{aligned}
 y_l &= h(x_l) + \mathcal{F}(x_l, \mathcal{W}_l), \\
 x_{l+1} &= f(y_l),
 \end{aligned}
 \tag{3}$$

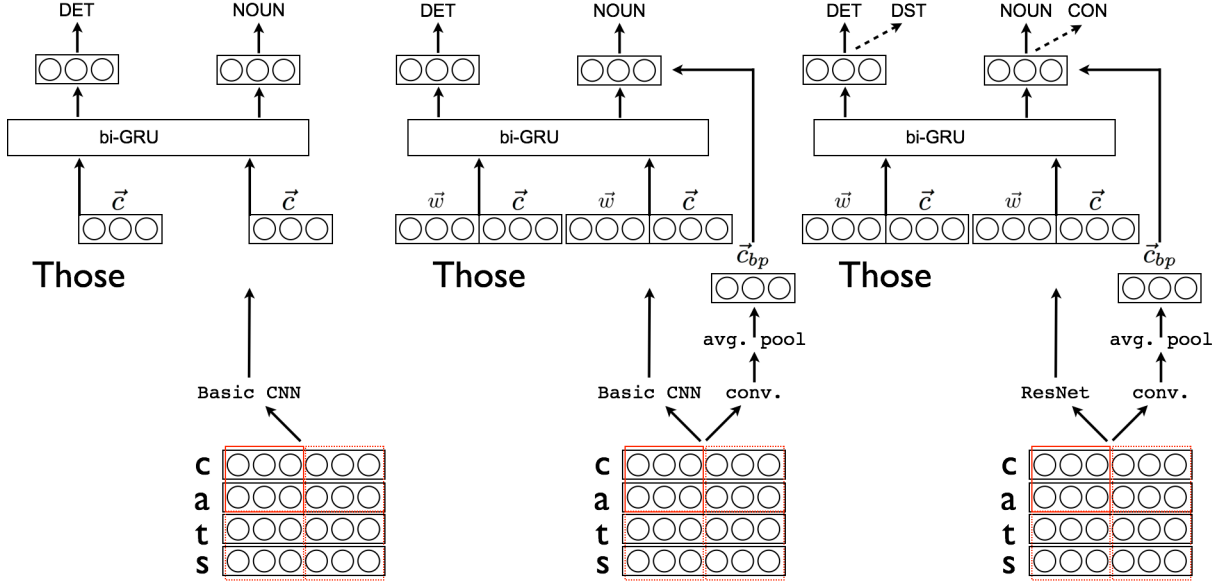


Figure 1: Model architecture. Left: Architecture with basic CNN char representations ( $\vec{c}$ ), Middle: basic CNN with char and word representations and bypass ( $\vec{c}_{bp} \wedge \vec{w}$ ), Right: ResNet with auxiliary loss and residual bypass (+AUX<sub>bp</sub>).

where  $x_l$  and  $x_{l+1}$  are the input and output of the  $l$ -th layer,  $\mathcal{W}_l$  is the weights for the  $l$ -th layer, and  $\mathcal{F}$  is a residual function (He et al., 2016), e.g., the identity function (He et al., 2015), which we also use in our experiments. ResNets can be intuitively understood by thinking of residual functions as paths through which information can propagate easily. This means that, in every layer, a ResNet learns more complex feature combinations, which it combines with the shallower representation from the previous layer. This architecture allows for the construction of much deeper networks. ResNets have recently been found to yield impressive performance in image recognition tasks, with networks as deep as 1001 layers (He et al., 2015; He et al., 2016), and are thus an interesting and effective alternative to simply stacking layers. In this paper we use the *asymmetric* variant of ResNets as described in Equation 9 in He et al. (2016):

$$x_{l+1} = x_l + \mathcal{F}(\hat{f}(x_l), \mathcal{W}_l). \quad (4)$$

ResNets have been very recently applied in NLP to morphological reinflection (Östling, 2016) and tasks such as sentiment analysis and text categorisation (Conneau et al., 2016). Our work is the first to apply ResNets to NLP sequence tagging tasks. We further contribute to the literature on ResNets by introducing a residual bypass function. The intuition is to combine both deep and shallow processing, which opens a path of easy signal propagation between lower and higher layers in the network.

### 3.3 Modelling character information and residual bypass

Using sub-token representations instead of, or in combination with, word-level representations has recently obtained a lot of attention due to their effectiveness (Sutskever et al., 2011; Chrupała, 2013; Zhang et al., 2015; Chung et al., 2016; Gillick et al., 2015). The use of sub-token representations can be approached in several ways. Plank et al. (2016) and Yang et al. (2016) use a hierarchical bi-directional RNN, first passing over characters in order to create word-level representations. Gillick et al. (2015) similarly apply an LSTM-based model using byte-level information directly. Dos Santos and Zadorozny (2014) construct character-based word-level representations by running a convolutional network over the character representations of each word. All of these approaches have in common that the character-based representation is passed through the entire remainder of the network. Our work is the

first to combine the use of character-level representations with both deep processing (i.e., passing this representation through the network) and shallow processing (i.e., bypassing the network in our residual bypass function). We achieve this by applying our novel residual bypass function to our character representations, inspired by the success of ResNets. In particular, we first apply the bypass to a CNN-based model achieving large gains over a plain CNN, and later evaluate its effectiveness in a ResNet.

A core intuition behind CNNs is the processing of an input signal in a hierarchical manner (LeCun et al., 1998; Goodfellow et al., 2016). Taking, e.g., a 3-dimensional image ( $width \times height \times depth$ ), the approach is typically to reduce spatial dimensions of the image while increasing depth. This hierarchical processing allows a CNN to learn high-level features of an input, essential to image recognition tasks. A drawback of this method, however, is that lower-level features are potentially lost in the abstraction to higher-level features. This issue is partially countered by ResNets, as information is allowed to flow more easily between residual blocks. However, this approach does not allow for simple and direct use of information in the network input in final layers. To alleviate this issue, we present a residual bypass function, which can be seen as a global residual function (depicted in Figure 1). This function allows both lower-level and higher-level features to be taken directly into account in the final layers of the network. The intuition behind using such a global residual function in NLP is that character information primarily ought to be of importance for the prediction of the current word. Hence, allowing these representations to bypass our bi-GRU might be beneficial. This residual bypass function is not dependent on the usage of ResNets, and can be combined with other NN architectures as in our experiments. We define the penultimate layer of a network with  $n$  layers, using a residual bypass, as follows:

$$y_{n-1} = h(x_{n-1}) + \mathcal{F}(x_i, \mathcal{W}_i), \quad (5)$$

where  $x_i$  and  $\mathcal{W}_i$  are the input and weights of the  $i_{th}$  layer,  $\mathcal{F}$  is a residual function (in our case the identity function), and  $h(x_{n-1})$  is the output of the penultimate layer. In our experiments, we apply a residual bypass function to our convolutional character representations.

### 3.4 System description

The core of our architecture consists of a bi-GRU taking an input based on words and/or characters, with an optional residual bypass as defined in subsection 3.3. We experiment with a basic CNN, ResNets and our novel residual bypass function. We also implemented a variant of the *Inception* model (Szegedy et al., 2015), but found this to be outperformed by ResNets. Our system is implemented in Keras using the Tensorflow backend (Chollet, 2015; Abadi et al., 2016), and the code is available at <https://github.com/bjerva/semtagger>.

We represent each sentence using both a character-based representation ( $S_c$ ) and a word-based representation ( $S_w$ ). The character-based representation is a 3-dimensional matrix  $S_c \in \mathbb{R}^{s \times w \times d_c}$ , where  $s$  is the zero-padded sentence length,  $w$  is the zero-padded word length, and  $d_c$  is the dimensionality of the character embeddings. The word-based representation is a 2-dimensional matrix  $S_w \in \mathbb{R}^{s \times d_w}$ , where  $s$  is the zero-padded sentence length and  $d_w$  is the dimensionality of the word embeddings. We use the English Polyglot embeddings (Al-Rfou et al., 2013) in order to initialise the word embedding layer, but also experiment with randomly initialised word embeddings.

Word embeddings are passed directly into a two-layer bi-GRU (Chung et al., 2014). We also experimented using a bi-LSTM. However, we found GRUs to yield comparatively better validation data performance on semtags. We also observe better validation data performance when running two consecutive forward and backward passes before concatenating the GRU layers, rather than concatenating after each forward/backward pass as is commonplace in NLP literature.

We use CNNs for character-level modelling. Our basic CNN is inspired by dos Santos and Zadrozny (2014), who use character-representations to produce local features around each character of a word, and combine these with a maximum pooling operation in order to create fixed-size character-level word embeddings. The convolutions used in this manner cover a few neighbouring letters at a time, as well as the entire character vector dimension ( $d_c$ ). In contrast to dos Santos and Zadrozny (2014), we treat a word analogously to an image. That is to say, we see a word of  $n$  characters embedded in a space

with dimensionality  $d_c$  as an image of dimensionality  $n \times d_c$ . This view gives us additional freedom in terms of sizes of convolutional patches used, which offers more computational flexibility than using only, e.g.,  $4 \times d_c$  convolutions. This view is applied to all CNN variations explored in this work.

A neural network is trained with respect to some loss function, such as the cross-entropy between the predicted tag probability distribution and the gold probability distribution. Recent work has shown that the addition of an auxiliary loss function can be beneficial to several tasks. Cheng et al. (2015) use a language modelling task as an auxiliary loss, as they attempt to predict the next token while performing named entity recognition. Plank et al. (2016) use the log frequency of the current token as an auxiliary loss function, and find this to improve POS tagging accuracy. Since our semantic tagging task is based on predicting fine semtags, which can be mapped to coarse semtags, we add the prediction of these coarse semtags as an auxiliary loss for the semtagging experiments. Similarly, we also experiment with POS tagging, where we use the fine semtags as an auxiliary information.

### 3.4.1 Hyperparameters

All hyperparameters are tuned with respect to loss on the semtag validation set. We use rectified linear units (ReLUs) for all activation functions (Nair and Hinton, 2010), and apply dropout with  $p = 0.1$  to both input weights and recurrent weights in the bi-GRU (Srivastava et al., 2014). In the CNNs, we apply batch normalisation (Ioffe and Szegedy, 2015) followed by dropout with  $p = 0.5$  after each layer. In our basic CNN, we apply a  $4 \times 8$  convolution, followed by  $2 \times 2$  maximum pooling, followed by  $4 \times 4$  convolution and another  $2 \times 2$  maximum pooling. Our ResNet has the same setup, with the addition of a residual connection. We also experimented with using average pooling instead of maximum pooling, but this yielded lower validation data performance on the semantic tagging task. We set both  $d_c$  and  $d_w$  to 64. All GRU layers have 100 hidden units. All experiments were run with early stopping monitoring validation set loss, using a maximum of 50 epochs. We use a batch size of 500. Optimisation is done using the ADAM algorithm (Kingma and Ba, 2014), with the categorical cross-entropy loss function as training objective. The main and auxiliary loss functions have a weighting parameter,  $\lambda$ . In our experiments, we weight the auxiliary loss with  $\lambda = 0.1$ , as set on the semtag auxiliary task.

Multi-word expressions (MWEs) are especially prominent in the semtag data, where they are annotated as single tokens. Pre-trained word embeddings are unlikely to include entries such as ‘International Organization for Migration’, so we apply a simple heuristic in order to avoid treating most MWEs as unknown words. In particular, the representation of a MWE is set to the sum of the individual embeddings of each constituent word.

## 4 Evaluation

We evaluate our tagger on two tasks: semantic tagging and POS tagging. Note that the tagger is developed solely on the semantic tagging task, using the GMB silver training and validation data. Hence, no further fine-tuning of hyperparameters for the POS tagging task is performed. We calculate significance using bootstrap resampling (Efron and Tibshirani, 1994). We manipulate the following independent variables in our experiments:

1. character and word representations ( $\vec{w}, \vec{c}$ );
2. residual bypass for character representations ( $\vec{c}_{bp}$ );
3. convolutional representations (Basic CNN and ResNets);
4. auxiliary loss (using coarse semtags on ST and fine semtags on UD).

We compare our results to four baselines:

1. the most frequent baseline per word (MFC), where we assign the most frequent tag for a word in the training data to that word in the test data, and unseen words get the global majority tag;
2. the trigram statistic based TNT tagger offers a slightly tougher baseline (Brants, 2000);

- the BI-LSTM baseline, running the off-the-shelf state-of-the-art POS tagger for the UD dataset (Plank et al., 2016) (using default parameters with pre-trained Polyglot embeddings);
- we use a baseline consisting of running our own system with only a BI-GRU using word representations ( $\vec{w}$ ), with pre-trained Polyglot embeddings.

#### 4.1 Experiments on semantic tagging

We evaluate our system on two semantic tagging (ST) datasets: our silver semtag dataset and our gold semtag dataset. For the +AUX condition we use coarse semtags as an auxiliary loss. Results from these experiments are shown in Table 3.

	BASELINES				BASIC CNN				RESNET					
	MFC	TNT	BI-LSTM	BI-GRU	$\vec{c}$	$\vec{c}_{bp}$	$\vec{c}_{bp} \wedge \vec{w}$	+AUX <sub>bp</sub>	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX	$\vec{c}_{bp}$	$\vec{c}_{bp} \wedge \vec{w}$	+AUX <sub>bp</sub>
ST Silver	84.64	92.09	94.98	94.26	91.39	90.18	94.63	94.53	94.39	95.14	94.23	94.23	<b>95.15</b>	94.58
ST Gold	77.39	80.73	82.96	80.26	69.21	65.77	76.83	80.73	76.89	<b>83.64</b>	74.84	75.84	82.18	73.73

Table 3: Experiment results on semtag (ST) test sets (% accuracy). MFC indicates the per-word most frequent class baseline, TNT indicates the TNT tagger, and BI-LSTM indicates the system by Plank et al. (2016). BI-GRU indicates the  $\vec{w}$  only baseline.  $\vec{w}$  indicates usage of word representations,  $\vec{c}$  indicates usage of character representations, and  $\vec{c}_{bp}$  indicates usage of residual bypass of character representations. The +AUX column indicates the usage of an auxiliary loss.

#### 4.2 Experiments on Part-of-Speech tagging

We evaluate our system on v1.2 and v1.3 of the English part of the Universal Dependencies (UD) data. We report results for POS tagging alone, comparing to commonly used baselines and prior work using LSTMs, as well as using the fine-grained semantic tags as auxiliary information. For the +AUX condition, we train a single joint model using a multi-task objective, with POS and ST as our two tasks. This model is trained on the concatenation of the ST silver data with the UD data, updating the loss of the respective task of an instance in each iteration. Hence, the weights leading to the UD softmax layer are not updated on the ST silver portion of the data, and vice-versa for the ST softmax layer on the UD portion of the data. Results from these experiments are shown in Table 4.

	BASELINES				BASIC CNN				RESNET					
	MFC	TNT	BI-LSTM	BI-GRU	$\vec{c}$	$\vec{c}_{bp}$	$\vec{c}_{bp} \wedge \vec{w}$	+AUX <sub>bp</sub>	$\vec{c}$	$\vec{c} \wedge \vec{w}$	+AUX	$\vec{c}_{bp}$	$\vec{c}_{bp} \wedge \vec{w}$	+AUX <sub>bp</sub>
UD v1.2	85.06	92.66	95.17	94.39	77.63	83.53	94.68	95.19	92.65	94.92	<b>95.71</b>	92.45	94.73	95.51
UD v1.3	85.07	92.69	95.04	94.32	77.51	82.89	94.89	95.34	92.63	94.88	<b>95.67</b>	92.86	94.69	95.57

Table 4: Experiment results on Universal Dependencies (UD) test sets (% accuracy). Adding semtags as auxiliary tags results in the best results obtained so far on English UD datasets.

## 5 Discussion

### 5.1 Performance on semantic tagging

The overall best system is the ResNet combining both word and character representations  $\vec{c} \wedge \vec{w}$ . It outperforms all baselines, including the recently proposed RNN-based bi-LSTM. On the ST silver data, a significant difference ( $p < 0.01$ ) is found when comparing our best system to the strongest baseline (BI-LSTM). On the ST gold data, we observe significant differences at the alpha values recommended by Søgaard et al. (2014), with  $p < 0.0025$ . The residual bypass effectively helps improve the performance of the basic CNN. However, the tagging accuracy of the CNN falls below baselines. In addition, the large gap between gold and silver data for the CNN shows that the CNN model is more prone to overfitting, thus favouring the use of the ResNet. Adding the coarse-grained semtags as auxiliary task only helps for

the weaker CNN model. The ResNet does not benefit from this additional information, which is already captured in the fine-grained labels.

It is especially noteworthy that the ResNet character-only system performs remarkably well, as it outperforms the BI-GRU and TNT baselines, and is considerably better than the basic CNN. Since performance increases further when adding in  $\vec{w}$ , it is clear that the character and word representations are complimentary in nature. The high results for characters only are particularly promising for multilingual language processing, a direction we want to explore next.

## 5.2 Performance on POS tagging

Our system was tuned solely on semtag data. This is reflected in, e.g., the fact that even though our  $\vec{c} \wedge \vec{w}$  ResNet system outperforms the Plank et al. (2016) system on semtags, we are substantially outperformed on UD 1.2 and 1.3 in this setup. However, adding an auxiliary loss based on our semtags markedly increases performance on POS tagging. In this setting, our tagger outperforms the BI-LSTM system, and results in new state-of-the-art results on both UD 1.2 (95.71% accuracy) and 1.3 (95.67% accuracy). The difference between the BI-LSTM system and our best system is significant at  $p < 0.0025$ .

The fact that the semantic tags improve POS tagging performance reflects two properties of semantic tags. Firstly, it indicates that the semantic tags carry important information for the prediction of POS tags. This should come as no surprise, considering the fact that the semtags abstract over and carry more information than POS tags. Secondly, it indicates that the new semantic tagset and released dataset are useful for downstream NLP tasks. In this paper we show this by using semtags as an auxiliary loss. In future work we aim to investigate the effect of introducing the semtags directly as features into the embedded input representation.

## 5.3 ResNets for sequence tagging

This work is the first to apply ResNets to NLP tagging tasks. Our experiments show that ResNets significantly outperform standard convolutional networks on both POS tagging and semtagging. ResNets allow better signal propagation and carry lower risk of overfitting, allowing for the model to capture more elaborate feature representations than in a standard CNN.

## 5.4 Pre-trained embeddings

In our main experiments, we initialised the word embedding layer with pre-trained polyglot embeddings. We also compared this with initialising this layer from a uniform distribution over the interval  $[-0.05, 0.05]$ . For semantic tagging, the difference with random initialisation is negligible, with pre-trained embeddings yielding an increase in about 0.04% accuracy. For POS tagging, however, using pre-trained embeddings increased accuracy by almost 3 percentage points for the ResNet.

## 6 Conclusions

We introduce a semantic tagset tailored for multilingual semantic parsing. We evaluate tagging performance using standard CNNs and the recently emerged ResNets. ResNets are more robust and result in our best model. Combining word and ResNet-based character representations helps to outperform state-of-the-art taggers on semantic tagging. Coupling this with an auxiliary loss from our semantic tagset yields state-of-the-art performance on the UD 1.2 and 1.3 POS datasets.

## Acknowledgements

The authors would like to thank Robert Östling for tips on ResNets, and Calle Börstell and Johan Sjons for feedback on earlier versions of this manuscript. We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. This work was partially funded by the NWO-VICI grant “Lost in Translation – Found in Meaning” (288-89-003).



## References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*.
- Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *ACL*, pages 1415–1425.
- Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. Forthcoming. The Groningen Meaning Bank. In Nancy Ide and James Pustejovsky, editors, *The Handbook of Linguistic Annotation*. Springer, Berlin.
- Johan Bos. 2008. Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.
- Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics.
- Alastair Butler. 2010. *The Semantics of Grammatical Dependencies*, volume 23. Emerald Group Publishing Limited.
- Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-domain name error detection using a multi-task rnn. In *EMNLP*.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.
- François Chollet. 2015. Keras. <https://github.com/fchollet/keras>.
- Grzegorz Chrupała. 2013. Text segmentation with character-level text embeddings. In *Workshop on Deep Learning for Audio, Speech and Language Processing, ICML*.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *Proceedings of ACL 2016, arXiv preprint arXiv:1603.06147*.
- Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.
- Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. 2005. Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, 3(2–3):281–332.
- James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic.
- Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.
- Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.
- Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426.

- Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2015. Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*.
- Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *arXiv preprint arXiv:1510.00726*.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep learning. <http://www.deeplearningbook.org>. Book in preparation for MIT Press.
- Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.
- Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL 2016*, *arXiv preprint arXiv:1604.05529*.
- Sameer S Pradhan, Wayne Ward, Kadri Hacioglu, James H Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240.
- Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martinez. 2014. Whats in a p-value in nlp? In *CoNLL-2014*.
- Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
- Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.
- Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215*.
- Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.